# BLIND AUDIO SOURCE SEPARATION OF STEREO MIXTURES USING BAYESIAN NON-NEGATIVE MATRIX FACTORIZATION

*S. Mirzaei*[*], *H. Van hamme*[*], *Y. Norouzi*[†]

[*] KU Leuven
ESAT Department
Leuven, Belgium

[†] Amirkabir University of technology
Department of Electrical Engineering
Tehran, Iran

## ABSTRACT

In this paper, a novel approach is proposed for estimating the number of sources and for source separation in convolutive audio stereo mixtures. First, an angular spectrum-based method is applied to count and locate the sources. A non-linear GCC-PHAT metric is exploited for this purpose. The estimated channel coefficients are then utilized to obtain a primary estimate of the source spectrograms through binary masking. Afterwards, the individual spectrograms are decomposed using a Bayesian NMF approach. This way, the number of components required for modeling each source is inferred based on data. These factors are then utilized as initial values for the EM algorithm which maximizes the joint likelihood of the 2-channel data to extract the individual source signals. It is shown that this initialization scheme can greatly improve the performance of the source separation over random initialization. The experiments are performed on synthetic mixtures of speech and music signals.

***Index Terms***— Blind Source Separation (BSS), Bayesian Non-negative Matrix Factorization(NMF), Marginal Maximum Likelihood (MML), Expectation-Maximization (EM)

## 1. INTRODUCTION

Demixing the audio signals has found many applications in several fields including polyphonic music source separation and transcription, speaker diarization and meeting transcription. Non-negative Matrix Factorization (NMF) has been applied extensively to various source separation scenarios in a single channel [1–3] and multichannel [4, 5] setting. In [5], a Non-Negative Tensor Factorization (NTF) structure is utilized for decomposing the magnitude spectrogram (Short Time Fourier Transform (STFT) representation) of the mixture signal into spectral components, time activations and channel mixing coefficients. The approach is suitable for instantaneous mixtures. In [4], the performance of the EM algorithm for maximizing the joint likelihood of multichannel data is explored. In [6], an extension of the complex matrix factorization method introduced in [7] is applied. The so called W-disjoint orthogonality of sources is presumed in [6]; This means that only one dominant source is assumed active in each time-frequency(TF) cell of the mixture signal STFT. This can be regarded as a sparsity assumption in the TF representation that might work for speech sources but is generally not valid for music.

The main disadvantage of the above methods is that the number of the sources as well as the number of the components used for modeling each source are assumed known and pre-defined. However, in practice, we may need to estimate them based on data. Here, we aim to overcome the mentioned drawbacks. Furthermore, in order to preserve the applicability of the algorithm to audio signals in general, our proposed source separation method doesn't rely on the sparsity assumption stated in [6]. In the first stage, we intend to estimate the number of the sources and channel mixing coefficients. This goal is achieved by evaluating an angular spectrum which is acquired from a non-linear Generalized Cross Correlation with Phase Transform (GCC-PHAT) metric calculated for individual TF cells of the spectrogram.

The task of source separation is addressed using the same EM method proposed in [4]. Our work differs from [4] in two aspects: First, we don't assume a pre-defined number of sources or model components and try to infer them based on the observed data. The second difference is related to our proposed initialization scheme which results in significant performance improvement. The primary estimated source power spectrograms derived from a binary masking technique are decomposed into two matrices containing the spectral components and time activations. A Bayesian NMF framework is exploited for this purpose which enables us to infer the number of components required to model each source. This is made possible by evaluating the Marginal Log-likelihood function against a range of model order values and finding the knee point. The Maximum Marginal Likelihood Estimation (MMLE) approach introduced in [8] is applied. A *Poisson-Gamma* generative model is assumed for the power spectrogram of the mixture signal. The efficiency of this Bayesian approach for inferring the optimal model order has already been investigated in [9]. Consequently, the obtained factors are used as initial values of the EM algorithm for extracting the individual source complex STFT representation. Time-

domain signals are easily derived through an inverse STFT operation.

The received mixture signal STFT based on the far-field assumption can be stated as follows:

$$X_{ift} = \sum_{j=1}^{J} S_{jft} a_{ijf} + n_{ift}\,, i = 1, 2 \tag{1}$$

where $X_{ift}$ denotes the complex value of the mixture signal STFT in frequency bin *f* and time frame *t* for $i^{th}$ channel. $n_{ift}$ can be representative of ambient noise or reverberation effects due to the room acoustics. $a_{ijf} = \exp(\frac{j2\pi df(i-1)\cos(\theta_j)}{c})$ is the channel mixing coefficient. *d* is the distance between the two microphones, *c* is the sound propagation velocity. $\theta_j$ denotes the angle of arrival for source *j* with respect to the microphone array axis. $S_{jft}$ is the complex contribution of each source in each TF bin and *J* is the total number of sources.

The rest of the paper is organized as follows: Section 2 is dedicated to the source counting and localization tasks. In section 3, the binary masking scheme and the proposed Bayesian NMF decomposition approach for inferring the components needed for modeling each source are explained. Afterwards, the EM framework is described in this section. The experiments and discussion on the results are presented in section 4. Section 5 concludes.

## 2. SOURCE COUNTING AND CHANNEL ESTIMATION

In order to estimate the angle of arrival of the source signals, we compute a metric against a range of angle of arrival(AOA) values $\theta$ aligned uniformly in the interval $[0, \pi]$. First, we evaluate the following function based on GCC-PHAT [10] in each TF bin against $\theta$ values:

$$R(f,t,\theta) = real(\frac{X_{1ft}X_{2ft}^*}{\left|X_{1ft}X_{2ft}^*\right|}\exp(\frac{j2\pi df\cos(\theta)}{c})) \tag{2}$$

where * denotes the conjugate operation. For increasing the spatial resolution, a monotonically decreasing non-linear function in the range $[0, 1]$ is applied to the GCC-PHAT metric based on what is proposed in [11]:

$$M(f,t,\theta) = 1 - \tanh(\alpha\sqrt{1 - R(f,t,\theta)}) \tag{3}$$

where $\alpha$ is the non-linearity parameter. This non-linear function makes the algorithm more effective in reverberant environments because it results in sharper peaks corresponding to the true source AOAs. To obtain the final angular spectrum, $F(\theta)$, a summation over all frequency bins and a maximization over all time frames is performed:

$$F(\theta) = \max_t \sum_f M(f,t,\theta) \tag{4}$$

To purify the angular spectrum, it can be obtained based on the TF cells which more probably correspond to one dominant active source. This way, the true peak levels corresponding to the actual source AOAs are enhanced and consequently diminish the effect of the spurious peaks. For identifying the mentioned cells, the following weighting function is introduced:

$$\lambda(f,t) = \begin{cases} 1 & \begin{cases} ||X_{1ft}| - |X_{2ft}|| < \gamma_1 \\ (|X_{1ft}| + |X_{2ft}|) > \gamma_2 \end{cases} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\gamma_2 = \max_t \left( \underset{f}{mean}(|X_{1ft}| + |X_{2ft}|) \right)$ and $\gamma_1$ is set to $10^{-5}$. The first condition in (5) is necessary for cells with one dominant source and the second one eliminates the contribution of the cells with relatively smaller magnitude. The metric expressed in (3) is then multiplied by this weighting to provide an improved angular spectrum.

Subsequently, a peak finding algorithm is applied to $F(\theta)$ to obtain the number of sources and AOAs. First the minimum value of the angular spectrum $F(\theta)$ is subtracted and then it is normalized, i.e. the vector is divided by its maximum value. Afterwards, two constraints on the minimum distance between the peaks and minimum peak height can eliminate irrelevant peak locations found by the peak finder algorithm. Here, we put the threshold for minimum peak height to 0.55 and for minimum peak distances to 5 degrees. These choices have led to optimum performance empirically even in reverberation conditions.

## 3. SOURCE SEPARATION

In this section, the estimated AOAs from the previous stage are utilized to extract the individual source signals. A primary estimation of each source complex spectrogram is constructed via binary masking. Then, a Bayesian MMLE approach is applied to each source spectrogram to infer the components required for modeling each source. The spectral components and time activations obtained in this stage along with the channel mixing coefficients given by the previous stage will construct the initial values of the parameters for the EM algorithm which accomplishes the source separation task.

### 3.1. Binary masking

Knowing the channel mixing coefficients, source separation can be done using TF masking techniques [12, 13]. For binary masking, we discriminate the dominant active source in each TF bin by evaluating the M function of (3) for the estimated AOA values found in the previous step and choosing

the source that maximizes this metric:

$$\begin{aligned} i_{BM}(f,t) &= \max_j M(f,t,\theta_j) \quad j = 1...J \\ S^{BM}_{i_{BM}(f,t)ft} &= X_{1ft} \\ S^{BM}_{jft} &= 0 \qquad \forall j \neq i_{BM}(f,t) \end{aligned} \tag{6}$$

where $i_{BM}(f,t)$ specifies the index of the recognized dominant source in frequency bin *f* and time frame *t*. $S^{BM}_{jft}$ denotes the complex spectrogram of the $j^{th}$ separated source.

### 3.2. Bayesian MMLE framework

The power spectrogram of the individual sources separated through binary masking are factorized according to the following Bayesian generative model:

$$\begin{aligned} v_{jft} &\sim \mathrm{P}oisson(v_{jft}|\sum_{k=1}^{K_j} w_{jfk}h_{jkt}) \\ h_{jkt} &\sim Gamma(h_{jkt}|\alpha_{jkt},\beta_{jkt}) \end{aligned} \tag{7}$$

where the involved distributions are defined by $\mathrm{P}oisson(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{\Gamma(x+1)}$ and $Gamma(x|\alpha,\beta) = [\beta^\alpha\Gamma(\alpha)]^{-1}x^{\alpha-1}e^{\frac{-x}{\beta}}$ respectively. $v_{jft}$ denotes the power spectrogram $\left|S^{BM}_{jft}\right|^2$, and $K_j$ is the total number of model components of the $j^{th}$ source. To impose sparsity, the elements of the activation matrix $\mathbf{H_j}$ are taken with *Gamma* prior distribution. To avoid overfitting, the Bayesian MMLE scheme was proposed in [8] which has been shown to automatically prune out irrelevant components (columns) of the $\mathbf{W_j}$ matrix, hence being capable of estimating the proper model order [9]. The elements of the $\mathbf{W_j}$ matrix containing the spectral components of the $j^{th}$ source are assumed deterministic. The log-likelihood is integrated over the $\mathbf{H_j}$ parameters and the obtained marginal log-likelihood to be maximized is:

$$C_{ML}(\mathbf{W}) = \log P(\mathbf{V}|W) = \log \int_{\mathbf{H}} P(\mathbf{V}|\mathbf{W},\mathbf{H})P(\mathbf{H})d\mathbf{H} \tag{8}$$

Since this integral is intractable, a lower bound on the marginal log-likelihood is maximized. For estimating the parameters, a variational Bayesian approach can be utilized. The update equations can be found in [8]. For finding the optimal model order, $K_j$, the lower bound of the marginal log-likelihood is evaluated against different order values. The knee point of the resulted graph specifies the proper model order [9]. This way, the number of components required for modeling each source and the components themselves are adaptively inferred based on binary masked separated sources.

### 3.3. Joint likelihood maximization using EM

The $\mathbf{W_j}$ and $\mathbf{H_j}$ components obtained in the previous subsection are applied as initial values of the EM algorithm in this stage. The initial value of the channel mixing coefficients is taken as $\underline{a}_{jf} = [1 \;\; \exp(\frac{j2\pi df\cos(\theta_j)}{c})]^T$ where $\theta_j$ denotes the estimated AOA of the $j^{th}$ source. The framework proposed in [4] is considered in which complex random variables $S_{jfn}$ of each source are assumed to be a sum of $K_j$ components:

$$S_{j,ft} = \sum_{k\in K_j} c_{k,ft} \quad c_{k,ft} \sim N_C(0, w_{fk}h_{kt}) \tag{9}$$

where $N_c$ is the proper complex Gaussian distribution. The components are assumed mutually independent and individually independent across frequency *f* and frame *t*, so we have [4]:

$$S_{j,ft} \sim N_C(0, \sum_{k\in K_j} w_{fk}h_{kt}) \tag{10}$$

A stationary and spatially uncorrelated noise model is presumed such that $n_{ift} \sim N_C(0,\sigma^2_{if}),\;\; \Sigma_{n,f} = diag([\sigma^2_{if}]_i)$. The set of all unknown parameters $\mathbf{Z} = \{\mathbf{A},\mathbf{W},\mathbf{H},\mathbf{\Sigma_n}\}$ including channel mixing coefficients, spectral components, time activations and noise covariance matrices are obtained by ML estimation which is solved via the EM algorithm. Iterative update relations can be found in [4]. The Complex STFT of the sources are ultimately obtained by the Minimum Mean Square Error (MMSE) estimate, $\widehat{S}_{jft} = E[S_{jft}|\mathbf{X_{ft}};\mathbf{Z}]$ where $\mathbf{X_{ft}}$ denotes the observed stereo mixture complex spectrogram.

The EM algorithm is very sensitive to parameter initialization. We have overcome this by introducing an effective initialization scheme. Furthermore, the proper number of components per source is chosen through an automatic model order selection using the Bayesian NMF model proposed in subsection 3.2.

## 4. EXPERIMENTS

The experiments are performed on synthetic stereo mixtures of speech and music signals taken from dev2 dataset of the SiSEC'08 "under-determined speech and music mixtures" task [14]. The sampling frequency is 16 kHz. The time duration of all individual sources is 10s. The STFT is computed with half-overlapping sine windows of length 1024. Three source signals are synthetically mixed based on the far-field model given in (1). Here, we consider noiseless condition. The source directions w.r.t the microphone array axis are taken [20° 60° 110°]. Microphone spacing is equal to 10 cm. The angular spectrum is calculated for 180 uniformly spaced angles in the range $[0,\pi]$. The non-linearity parameter $\alpha$ is taken equal to 10. Applying the constraints to the peak finder algorithm lead to the estimated AOA values perfectly matched with the true ones.

The primary estimation of the individual source complex spectrograms is given by applying the binary masking technique. Then, the Bayesian NMF method is utilized to factorize the obtained power spectrograms. The variational

Bayesian algorithm is executed for the order values between 1 and 20 and the lower bound of marginal log-likelihood is plotted against model order. Then, the number of components required for modeling each source is implied through finding the knee point of this graph. The number of iterations for the variational inference algorithm is set to 1000. The impact of local optima was alleviated by executing the algorithm 10 times and choosing the results corresponding to the largest likelihood lower bound. For Bayesian NMF, the initial values of the $\mathbf{W_j}$ elements are taken as absolute value of a random normal variable (with mean 0 and variance 1) plus 1. Initial values of the shape, $\alpha_{jkt}$ and scale, $\beta_{jkt}$ hyperparameters of the $\mathbf{H_j}$ matrix prior distribution are set to 1. The log-likelihood lower bound is depicted in Figure 1 for the wdrums sources. To avoid an *ad hoc* method to determine the knee point of the graph, we exploit Bayesian Information Criterion (BIC) metric defined as:

$$BIC = -2LB + N_P \log(N_{obs}) \tag{11}$$

where *LB* represents the log-likelihood lower bound, $N_P$ denotes the number of model parameters and $N_{obs}$ is the number of observed samples associated with each value of LB. The order value corresponding to the minimum obtained BIC metric is taken as the optimal number of model components. The inferred orders for wdrums sources are 4, 9 and 5 respectively. The same task is carried on for the nodrums and the male speech sources. The estimated optimal order values are 7,4,8 and 10,11,12 respectively. It can be seen that more components are needed to model speech sources.

1000 iterations of the EM algorithm are then executed for extracting the final complex spectrogram of each source. The simulated annealing strategy proposed in [4] is applied for updating $\sigma_{if}^2$ in each iteration. Commonly used performance measures including Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifact Ratio (SAR) are calculated [15]. The results are listed in Tables 1-3 for speech and music mixtures. The performance improvement through the proposed initialization scheme is manifested especially for music mixtures. For random initialization, the number of model components per source is taken equal to 10. The initial values of the elements of the factors **W** and **H** in this case, are taken as absolute value of a random normal variable (with mean 0 and variance 1) plus 1. The significant lower performance achieved by random initialization for music signals probably originates from overfitting or local optima issues.

## 5. CONCLUSION

In this paper, a three stage approach was introduced for separating the sources in a stereo convolutive mixture scenario. In the first stage, the number of sources and the channel mixing coefficients are estimated by finding the peak locations of an

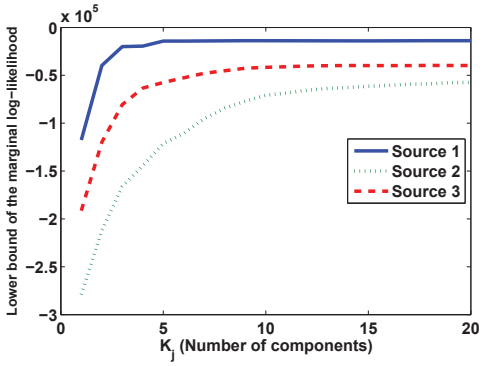


**Fig. 1**. Log-likelihood lower bound evaluated vs. model order for wdrums sources

| Male speech signal | | | |
|---|---|---|---|
| | S1 | S2 | S3 |
| | Proposed initialization | | |
| SDR(dB) | 8.92 | 12.1 | 16.41 |
| SIR(dB) | 15.6 | 23.5 | 23.64 |
| SAR(dB) | 10.1 | 12.45 | 17.33 |
| | Random initialization | | |
| SDR(dB) | 5.93 | 8.27 | 13.56 |
| SIR(dB) | 14.76 | 12.31 | 16.32 |
| SAR(dB) | 6.85 | 11.48 | 17.1 |

**Table 1**. BSS evaluation metrics for male speech data.

| nodrums | | | |
|---|---|---|---|
| | S1 | S2 | S3 |
| | Proposed initialization | | |
| SDR(dB) | 12.4 | 8.4 | 6.19 |
| SIR(dB) | 15.52 | 19.16 | 13.9 |
| SAR(dB) | 15.4 | 8.83 | 7.21 |
| | Random initialization | | |
| SDR(dB) | 0.96 | -1.08 | -8.14 |
| SIR(dB) | 7.43 | 0.53 | -7.22 |
| SAR(dB) | 2.7 | 6.73 | 7.03 |

**Table 2**. BSS evaluation metrics for nodrums data.

| wdrums | | | |
|---|---|---|---|
| | S1 | S2 | S3 |
| | Proposed initialization | | |
| SDR(dB) | 1.26 | 6.87 | 25.34 |
| SIR(dB) | 2.69 | 19.36 | 32.72 |
| SAR(dB) | 10.64 | 9.05 | 33.13 |
| | Random initialization | | |
| SDR(dB) | -5.73 | 1.39 | 11.8 |
| SIR(dB) | -4.5 | 9.95 | 12.73 |
| SAR(dB) | 5.18 | 2.45 | 19.22 |

**Table 3**. BSS evaluation metrics for wdrums data.

angular spectrum derived from non-linear GCC-PHAT metric. The mixing coefficients are then exploited in the second stage to obtain binary masks separating the complex spectrogram of the sources. The primary source power spectrograms given by binary masking are then individually decomposed through a Bayesian NMF approach, thus the number of components required for modeling each source is estimated adaptively and is not assumed pre-defined. Automatic model order selection accomplished in this stage is of great importance because it avoids an overfitted or underfitted model. For instance, speech signals generally need more components than music sources.

The decomposed factors are then employed as initial values in the third stage (EM algorithm). It has been shown that this initialization scheme can enhance the performance drastically compared to the random initialization. By applying the proposed effective initialization approach, the high sensitivity of the EM algorithm to parameter initialization is alleviated. This is demonstrated by evaluating the BSS metrics after applying the proposed method to the synthetic convolutive mixtures of speech and music.

## REFERENCES


[1] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, Oct 2003, pp. 177–180.

[2] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[3] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and Scott Rickard, "Clustering nmf basis functions using shifted nmf for monaural sound source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 245–248.

[4] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 550–563, March 2010.

[5] C. Fvotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues," in *Proc. 7th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, Mlaga, Spain, 2010, vol. 6684, pp. 102–115, Springer.

[6] K. Takeda, H. Kameoka, H. Sawada, S. Araki, S. Miyabe, T. Yamada, and S. Makino, "Underdetermined bss with multichannel complex nmf assuming w-disjoint orthogonality of source," in *TENCON 2011 - 2011 IEEE Region 10 Conference*, Nov 2011, pp. 413–416.

[7] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex nmf: A new sparse representation for acoustic signals," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3437–3440.

[8] O. Dikmen and C. Fevotte, "Maximum marginal likelihood estimation for nonnegative dictionary learning in the gamma-poisson model," *Signal Processing, IEEE Transactions on*, vol. 60, no. 10, pp. 5163–5175, Oct 2012.

[9] S. Mirzaei, H. Van hamme, and Y. Norouzi, "Model order estimation using bayesian nmf for discovering phone patterns in spoken utterances," in *Proc. Interspeech 2013*, August 2013.

[10] C. Knapp and G Clifford Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[11] L. Benedikt and B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in *Latent Variable Analysis and Signal Separation*, pp. 1–8. Springer, 2010.

[12] O. Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[13] V.G. Reju, Soo Ngee Koh, and I.Y. Soon, "Underdetermined convolutive blind source separation via time-frequency masking," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 1, pp. 101–116, Jan 2010.

[14] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brazil, 2009, pp. 734–741.

[15] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, July 2006.